

# A cumulative approach to quantification for sentiment analysis

Giambattista Amati, Simone Angelini, Marco Bianchi,  
Luca Costantini, and Giuseppe Marcone

Fondazione Ugo Bordoni, Viale del Policlinico 147, 00161 Roma

**Abstract.** *We estimate sentiment categories proportions for retrieval within large retrieval sets. In general, estimates are produced by counting the classification outcomes and then by adjusting such category sizes taking into account misclassification error matrix. However, both the accuracy of the classifier and the precision of the retrieval produce a large number of errors that makes difficult the application of an aggregative approach to sentiment analysis as a reliable and efficient estimation of proportions for sentiment categories. The challenge for real time analytics during retrieval is thus to overcome misclassification errors, and more importantly, to apply sentiment classification or any other similar post-processing analytics at retrieval time. We present a non-aggregative approach that can be applied to very large retrieval sets of queries.*

**Key words:** Information Retrieval, Sentiment Analysis, Quantification

## 1 Introduction

We study the problem of estimating the size and proportions of sentiment categories (category quantification [5,6,2,3]) over a result set of a query. The quantification problem is very challenging because of several factors: the sentiment content drift caused by the content of a query, the size of the result set, the term sparsity, the precision of the retrieval, finally the accuracy of the classifier. In sentiment quantification the number of classification errors (false positives and false negatives) as well as for each result set the ability of the classifier to balance the priors of the sentiment categories are both important. Indeed, existent test sets show that both error rates and categories priors may largely vary each topic or result set. There are four possible approaches to quantification: to adjust counts with a confusion matrix, to choose a suitable training set to learn the classifier in order to better fit the priors to the new data, to improve the quantification accuracy with a proper multivariate classification model, to smooth the classification counts with a second learning model.

The first approach (the AC&C approach) classifies documents in the retrieval set  $D_q$  over a certain number of categories  $c$ , and then counting of the elements (the set  $\hat{c}$ ) falling into each category  $c$  is eventually adjusted to the final estimate  $\hat{c} \cap D_q$  with the numbers of the misclassification errors that the classifier makes on a training set and that is provided by the confusion matrix  $p(\hat{c}_i|c_j)_{i \neq j}$  [5,6].

Among these approaches the empirical median sweep find exhaustively any possible classification threshold to obtain an estimate of the prevalence. The final prevalence quantity is the median of all the estimates.

The second approach is to use a set of spanning features, that must be drawn randomly and independently from the categories, that is then used to draw a suitable training sample from a validation set. Not all the validation set is used to train the classifier but a proper subset. The drawn training set turns out to be the closest set to the collection according to a distance, for example the Kullback-Leibler Divergence or the Hellinger distance [8]. Such distance is between the two distributions of the features: the first on the collection, the second on the retrieval set. Though the Hopkins and King method is not automatic [9], it can be still fall into such an approach, since it smooths the raw estimates of a manual evaluation by counting categories over a spanning set of features in the collection.

Since quantification accuracy is related to the ability of the quantification model to minimize the difference in size of false positives and false negatives, at certain extent the classification accuracy is independent from the quantification accuracy. However, it is also true that the higher the classifier accuracy is, the less the difference in size of the errors is, all other experiment settings remaining the same. One obstacle to achieve a higher quantification accuracy is that some classifiers are binary in nature (such as SVM or the approach based on the Hellinger Distance) so a different approach has shown to achieve a better quantification accuracy under a multivariate approach [10]. With a higher accuracy the multivariate approach avoids smoothing methods based on the confusion matrix.

The last approach is non aggregative and use two distinct learning models: the first is the classifier the second model learns how to quantify from the classifier. Instead of using the confusion matrix this approach does not use the classifier as a Bayesian decision rule but cumulates the scores used to emit such decisions and correlates observed categories sizes to such aggregate scores through for example regression models [1].

For particular dataset, such as the Internet Movie Dataset, there is also a link-based quantification model [4], and an iterative method [20] to rectify the classifier when the prevalence of a class may change over time. The Expectation Maximization can be also applied for adjusting the outputs of a classifier with new class priors[12].

We introduce a non aggregative approach on Sections 3 and 4. We define the experimental settings and the evaluation measures suitable in a retrieval scenario. In particular we use the Kolmogorov-Smirnov distance, and its p-value also provides a statistical significance test for validating the goodness-of-fitting of the new model.

## 2 Related Works

According to the family of the Adjusted Classify & Count methods, once the classifier returns a set  $\hat{c}$  for each category  $c$  in a proportion  $P(\hat{c}_j|q)$  among the  $n$  categories, the *Theorem of Total Probability* decomposes these classifier outcomes over the set of  $n$  categories [14]  $[P(\hat{c}_j|q) = \sum_{i=1}^n P(\hat{c}_j|c_i, q)P(c_i|q) \ j=1, \dots, n]$ . The Scaled Probability Average approach is a variant of the ACC method, with the expectation over the categories probabilities used instead of the total probability theorem[3]. The unknown estimates  $P(c_i|q)$  are derived solving a simple linear system of  $n$  equations with  $n$  variables:  $\underset{n \times n}{P(\hat{c}|c, q)} \cdot \underset{n \times 1}{P(c|q)} = \underset{n \times 1}{P(\hat{c}|q)}$ . The accuracy of the classifier should not matter, since the misclassification errors are used to estimate all category sizes. This model can be easily extended with a linear regression model to learn from a set of queries (or different training sets), i.e.  $\underset{n \times n}{P(\hat{c}|c, q)} \cdot \underset{n \times Q}{P(c|q)} \sim \underset{n \times Q}{P(\hat{c}|q)}$  or with an entropy value  $H$  substituted for  $P[1]$ ,  $\underset{n \times n}{H(\hat{c}|c, q)} \cdot \underset{n \times Q}{P(c|q)} \sim \underset{n \times Q}{P(\hat{c}|q)}$ . Here,  $\sim$  stands for equality up to linear regression coefficients that fit the the model with  $|Q|$  equations.

## 3 Cumulative Classifiers

We use the learning models of three classifiers, MNB, SVM and DBM, and apply a cumulative measure

$$\mu_c(\sum_d x_i | x_i \text{ frequency of } i \text{ in } d, d \in D_q) \quad (1)$$

for the retrieval set  $D_q$  and category  $c$  of documents, that is a measure satisfying the following property:  $\mu_c(\sum_i X_i) = \sum_{d \in D, i} \mu_c(x_i)$  with  $X_i = \sum_{d \in D} x_i$  and  $\sum_i \mu_c(x_i)$  used to classify documents  $\mathbf{x}$ . Such an additive property derived from a classifier, allows us to make the hypothesis that the cumulative function  $\mu_c(D_q)$  correlates (linearly) with the number of documents that are relevant to the query ( $\mathbf{x} \in R_q$ ) and are in the category  $c$ :

$$\Phi_c(\{\mu_c\}_{c \in \mathcal{C}}, D_q, \theta) = |R_q \cap c| \quad (2)$$

Obviously not all classifiers are suitable for defining such a cumulative measure, but MNB, DBM and SVM are.

Since the learning probabilistic model of MNB is based on the term independence assumption, the logarithm of probabilities is additive over terms. The Kullback-Leibler divergence is also additive [11], so that both the probabilistic learning model of DBM and MNB satisfy the additivity property over independent terms. Analogously, SVM can be seen as cumulative measure function with respect to the direction of the hyperplanes because distance is additive along that direction. We now show that these additive properties are necessary conditions in order to derive a cumulative measure for these three classifiers. Moreover, Table 1 shows that there is a linear correlation between the cumulative measures

of MNB and DBM over the categories and the cardinalities of their respective categories (positive versus negative).

**Multinomial Naive Bayes (MNB).** Due to the sparsity of data, Naive Bayes (NB) and gaussian Naive Bayes perform poorly in text classification[17], therefore the MNB classifier is preferred to NB. Let  $x_i$  be the frequency of word  $i$  in document  $d$ , and  $p(c)$  be the prior for category  $c$ , that is the frequency of category  $c$  in the training collection, and  $f_{i,c}$  the frequency of the word  $i$  in category  $c$  containing  $L_c$  tokens of words. Most of the implementations of MNB [16] maximize the logarithm of the likelihood with a multinomial distribution (one for each category) as follows:

$$\arg \max_c \left[ \log p(c) + \sum_i x_i \log \left( \frac{f_{i,c} + \alpha_i}{L_c + \alpha} \right) \right] \quad (3)$$

where  $\alpha = \sum_i \alpha_i$  and  $\alpha_i$  smoothing parameters. We choose  $\alpha_i = 1$  [17]. The cumulative function is

$$\mu_c(\mathbf{x}) = \sum_i x_i \log \left( \frac{f_{i,c} + \alpha_i}{L_c + \alpha} \right) \quad (4)$$

It is easy to verify that  $\mu_c(D) = \sum_{\mathbf{x} \in D} \mu_c(\mathbf{x})$ .

**SVM classifier.** SVM constructs a direction  $\mathbf{w} = \sum_j \alpha^j y^j \mathbf{x}^j$ ,  $\mathbf{x}^j$  being the vector containing all the frequencies of the  $j$ -th support document. The distances of documents  $\mathbf{x}$ , considered as vectors of terms, from the hyperplanes of equations  $\mathbf{w} \cdot \mathbf{x} + b = 1$  and  $\mathbf{w} \cdot \mathbf{x} + b = -1$  define a decision rule to assign a document to a category. Differently from probabilistic learning models, where we can exploit the additivity property of the logarithm function over independent events, we here make use of the additivity property of distance along the normal direction  $\mathbf{w}$  to both category hyperplanes. Then, we assume that the sum of the distances of documents from an hyperplane, that is  $\sum_{\mathbf{x} \in D} \mu_c(\mathbf{x})$ , is linearly correlated to the number  $|c|$  of the elements in the corresponding category, that is such an assumption is  $\sum_{\mathbf{x} \in D} \mu_c(\mathbf{x}) \propto |c|$ . Therefore, if  $\mu_c(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$ , then the sum  $\sum_{\mathbf{x} \in D} \mu_c(\mathbf{x}) = \sum_{\mathbf{x} \in D} (\mathbf{w} \cdot \mathbf{x} + b)$  with the constraint  $(\mathbf{w} \cdot \mathbf{x} + b) > 1$  (or  $\mathbf{w} \cdot \mathbf{x} + b < -1$  respectively) provides a number  $|\hat{c}|$  of positive (negative) documents in the result set  $D$ . We note that  $\mu_c(D) = \sum_{\mathbf{x} \in D} \mu_c(\mathbf{x})$  up to the additive constant  $b \cdot |\hat{c}|$ . The distributive property of inner product with respect to the sum of vectors  $\mathbf{x}$  implies the cumulative property of  $\mu_c$  up to an additive constant proportional to  $|\hat{c}|$ . However, the hypothesis of correlation between the cumulative function  $\mu_c(D)$  and the estimated number  $|\hat{c}|$  of elements of the category is not affected because  $\mu_c(D) \propto |\hat{c}|$  is equivalent to  $\mu_c(D) + b \cdot |\hat{c}| \propto |\hat{c}|$ .

**Divergence-Based Model, DBM.** The divergence based model DBM is a variant of the MNB of Equation 5. The difference with MNB (Equation 5) lies in considering the factorials part in the calculation of the likelihood probabilities [18]. The multinomial distribution with a prior probability distribution  $\pi_i$  over  $L$  tokens can be approximated using the Kullback Leibler divergence. If the frequency of the term  $i$  in a category  $c$  equals  $f_{i,c} = \left\{ \frac{\sum_{\mathbf{x} \in c} x_{i,c}}{L} \right\}$  when  $x_{i,c}$  is

the frequency of the term  $i$  in document  $d \in c$ , then an approximation of the multinomial is:

$$\sum_i f_{i,c} \cdot \log \left( \frac{f_{i,c}}{\pi_i} \right) \quad (5)$$

leading to the decision rule for a document  $\mathbf{x}$ :

$$\arg \max_c \left[ \log p(c) + \sum_i x_i \cdot f_{i,c} \cdot \log \left( \frac{f_{i,c}}{\pi_i} \right) \right] \quad (6)$$

Note that, if  $l = \sum_i x_i$ ,  $\sum_i f_{i,c} \cdot \log \left( \frac{f_{i,c}}{\pi_i} \right) = l \cdot D(f_c | \pi)$  with  $D(f_c | \pi)$  the Kullback-Leibler divergence between the distributions  $\{f_{i,c}\}$  and  $\{\pi_i\}$ . Each token of the term  $i$  contributes with Formula (5) in the decision rule. The model is learned on a training sample of  $L$  tokens, drawing the frequencies  $f_{i,c}$ . The cumulative function is

$$\mu_c(\mathbf{x}) = \sum_i x_i \cdot f_{i,c} \cdot \log \left( \frac{f_{i,c}}{\pi_i} \right) \quad (7)$$

It is easy to verify that  $\mu_c(D) = \sum_{\mathbf{x} \in D} \mu_c(\mathbf{x})$ .

## 4 Cumulative Quantification Models

The cumulative approach requires a learning model  $\Phi$  to correlate the cumulative function of a classifier  $\mu_c$  with the category size  $c$  as shown on Formula 2. Essential parameters of  $\theta$  are the size of  $D_q$ , the sparsity of the terms, that is correlated to the size of the lexicon, that in turns grows following Heap's Law as long as new documents are indexed [13,15]. The size of the lexicon has a direct effect on the accuracy of the classifier. In a practical retrieval scenario one should expect to search, classify and count even millions of documents in one single shot so we need to verify whether quantification approaches can scale both in effectiveness and efficiency.

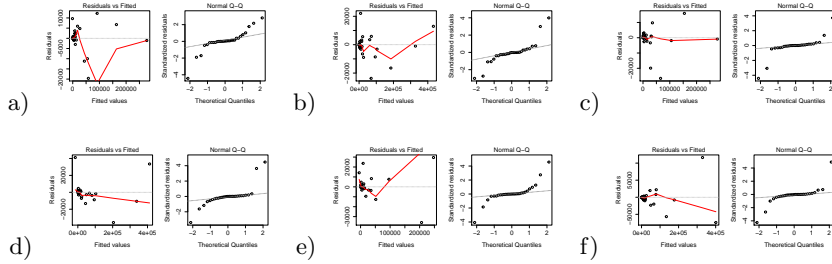
We first observe that there exists a strong *linear* correlation factor between cumulative sentiment  $\mu_c$  and category size. This hypothesis is statistically significant, as shown on Table 1. In force of this evidence, we may initially set  $\Phi$  to a linear regression model. However, we encounter the following problems:

a) Cumulative DBM and MNB learning models provide estimates of each category size that are learned independently from each other. Cumulative SVM instead uses both the two mutually exclusive categories in the training set, and the quantification methods based on the Hellinger distance [8] are only suitable for binary classification. In actual situations there are at least six categories (non-relevant, positive, negative, mixed, neutral and other). Since they are mutually exclusive they should satisfy the constraint that  $\sum_c \mu_c(D) = |D|$ . Some

of these categories are too difficult to be learned (such as the category of non-relevant document, the neutral or the “other” category). Moreover many documents contain both opposite sentiment polarities, therefore the mixed category  $M$  is mutually exclusive but difficult to separate in practise from positive and negatives ones,  $P$  and  $N$ . To address multiple categories classification, the multivariate SVM can be used instead of the confusion matrix. The training set is passed as one single input to the multivariate SVM and processed in one single pass. According to [7], the accuracy of multivariate SVM with a counting and classify approach shows a better performance than the adjusted counting and classify with confusion matrix and other variants.

b) The learning model for  $\Phi$  can be trained either by pooling all examples irrespective of the set of queries used for the retrieval evaluation (item-driven learning), or can be trained by aggregating results query-by-query (query-driven learning). In practise we train  $\Phi$  with an equation either for each observation or for each query.

c) In order to normalize categories one can include the size of the result set  $D_q$  as a parameter of a regression model. However, the regression model has several outliers, that occur when their result set is very large  $D_q$ . d) In real applications there exists a high variability of the categories priors. Besides the variability of the priors, the size of test data is very small in all available data sets. In conclusion, there always exists a mismatch between the training data distributions and the actual data distributions to which apply the quantification learning model.



**Fig. 1.** The regression parameters are learned by a query-driven experiments, that is setting one equation for each query  $q \in Q$ . **DBM** [a) b)], **MNB** [c) d)], **SVM** [e) f)].  
a), c), e)  $P + M \sim \alpha_P \cdot \mu_P(D_q) + \alpha_N \cdot \mu_N(D_q) + \beta_P \cdot |D_q|$  and  $q \in Q$ .  
b), d), f)  $N + M \sim \alpha'_P \cdot \mu_P(D_q) + \alpha'_N \cdot \mu_N(D_q) + \beta_N \cdot |D_q|$  and  $q \in Q$ .

## 5 Linear Regression Model

Cumulative quantification models have two sets of parameters: the parameters of the classifier (e.g. the support vectors and the parameter  $b$  for SVM,  $L_c$ ,

**Table 1.** Pearson Correlation  $\rho$  between cumulative measures and observed cardinalities for a set of queries for each classifier. The correlation is made on a leave-one-out cross validation learning observations. For each category  $c \in \{P, N\}$  the number of the two categories in the collection (estimated by sampling randomly in each result set of the queries and manually evaluated) is correlated to the cumulative measure of the classifiers  $\mu_c$  over the collection. We also report the confidence interval for  $\rho$  at 95% of confidence level.

classifier/ $\rho$	DBM		SVM		MNB	
	mean	interval	mean	interval	mean	interval
$\mu_P/P$ (estim.)	0.974	$\rho \in [0.944, 0.988]$	0.946	$\rho \in [0.887, 0.975]$	0.924	$\rho \in [0.887, 0.975]$
$\mu_N/N$ (estim.)	0.987	$\rho \in [0.973, 0.994]$	0.918	$\rho \in [0.831, 0.961]$	0.950	$\rho \in [0.895, 0.977]$

$L$ ,  $f_{i,c}$  and  $\pi_i$  for DBM and MNB) and the parameters of the learning model that correlates the cumulative measure of the classifier to the counting measure of categories size. We expect that the relation between the two measures is expressed by a linear correlation, therefore we choose the linear regression as natural learning model to express such a correlation (see Table 1).

We assume that we have already learned the classifier generating thus a cumulative measure  $\mu_c$  for each category  $c \in \mathcal{C}$ . We now consider a second learning model  $\Phi$  that learns how to predict category size and proportions from the cumulative measures  $\mu_c$ . The validation data set  $\Omega$  is made of about 3 million tweets. In order to validate the quantification model  $\Phi$  we exclude all relevant and evaluated tweets of a query  $q$  both to learn  $\mu$  and to predict the category sizes for the retrieval set of that query. The set of evaluated tweets  $V$  is extracted from a proper subset of  $\Omega$ :  $V = \cup_{q \in Q} V_q$  with  $V_q \subset R_q$  where  $R_q$  is the set of relevant documents that fall into 5 mutually exclusive categories:

$$R_q = [P_q \cup N_q \cup M_q \cup X_q \cup O_q]$$

**Table 2.** Pearson Correlation  $\rho$  between proportions with the classify and count models (CC( $c$ )), cumulative models  $\Phi$  and proportions of  $c$  for the set of queries. The correlation is made on a leave-one-out cross validation learning observations. We also report the confidence interval for  $\rho$  at 95% of confidence level (all p-values are  $< 0.05$ ).

	Classify and Count			Query driven $\Phi$			Item driven $\Phi$		
	DBM	MNB	SVM	DBM	MNB	SVM	DBM	MNB	SVM
Pos $\rho$ mean	0.886	0.943	0.951	0.991	0.989	0.978	0.978	0.975	0.943
Neg $\rho$ mean	0.979	0.954	0.943	0.994	0.988	0.936	0.99	0.986	0.928
Pos $\rho$ Inf	0.769	0.881	0.896	0.981	0.977	0.953	0.954	0.946	0.88
Neg $\rho$ Inf	0.955	0.902	0.88	0.987	0.974	0.866	0.977	0.97	0.85
Pos $\rho$ Sup	0.945	0.973	0.976	0.996	0.995	0.989	0.989	0.988	0.973
Neg $\rho$ Sup	0.99	0.978	0.973	0.997	0.994	0.969	0.995	0.993	0.965

**Table 3.** Kolmogorov-Smirnov distance  $D \in [0, 1]$  between the fitted and the observed distributions. The two sets of fitted and observed values come from the same distribution is the hypothesis under test. The \* indicates statistical significance at 95% level of confidence.

		Classify and Count					
		DBM		MNB		SVM	
		D	p-value	D	p-value	D	p-value
Pos		0.069	1 *	0.2069	0.5722	0.2759	0.2221
Neg		0.2414	0.372	0.2414	0.3669	0.2069	0.5722

		$\Phi$ Query driven						$\Phi$ Item driven					
		DBM		MNB		SVM		DBM		MNB		SVM	
		D	p-value	D	p-value	D	p-value	D	p-value	D	p-value	D	p-value
Pos		0.103	0.998 *	0.172	0.791	0.137	0.951 *	0.620	$1.5e^{-05}$	0.137	0.951 *	0.137	0.951 *
Neg		0.137	0.951 *	0.103	0.998 *	0.172	0.791	0.413	0.013	0.137	0.951 *	0.103	0.998 *

The category rates are estimated using the set of evaluated tweets  $V_q$ , i.e.:  $\hat{c}_q^\% = \frac{R_q \cap c_q \cap V_q}{R_q \cap V_q} = \frac{c_q \cap V_q}{V_q}$ , with  $c \in \{P, N, M, X, O\}$ . The actual values  $c_q^\% = \frac{c_q \cap R_q}{R_q}$  fall into a confidence interval, i.e.  $c_q^\% = \hat{c}_q^\% \pm \epsilon$ .

The linear regression models of the query-driven approach are learned by using the values  $\mu_P(D_q)$  and  $\mu_N(D_q)$  computed on the entire result set  $D_q$ :

$$\hat{P}_q^\% + M_q^\% \sim \alpha_P \cdot \frac{\mu_P(D_q)}{\sum_{c \in \{P, N\}} \mu_c(D_q)} + \alpha_N \cdot \frac{\mu_P(D_q)}{\sum_{c \in \{P, N\}} \mu_c(D_q)} \text{ s.t. } q \in Q \quad (8)$$

$$\hat{N}_q + M_q^\% \sim \alpha'_P \cdot \frac{\mu_P(D_q)}{\sum_{c \in \{P, N\}} \mu_c(D_q)} + \alpha'_N \cdot \frac{\mu_P(D_q)}{\sum_{c \in \{P, N\}} \mu_c(D_q)} \text{ s.t. } q \in Q \quad (9)$$

The number of positive documents  $P$  is estimated by  $\hat{P}_q = |D_q| \cdot \hat{P}_q^\%$ , and similarly with the set of negative documents  $\hat{N}_q = |D_q| \cdot \hat{N}_q^\%$ , where  $\hat{P}_q^\%$  and  $\hat{N}_q^\%$  are evaluated on the set  $V_q$ , and  $D_q$  is the retrieval set for the query  $q$ .

Differently, the item-driven approach learns the regression parameters by means of the following set of equations:

$$p(\mathbf{x}) \sim \alpha_P \cdot \mu_P(\mathbf{x}) + \alpha_N \cdot \mu_P(\mathbf{x}) \text{ if } \mathbf{x} \in P \cup N \quad (10)$$

$$n(\mathbf{x}) \sim \alpha'_P \cdot \mu_P(\mathbf{x}) + \alpha'_N \cdot \mu_P(\mathbf{x}) \text{ if } \mathbf{x} \in P \cup N \quad (11)$$

where  $p(\mathbf{x})$  is equal to 1 if the document  $\mathbf{x}$  is positive and 0 if it is negative; while,  $n(\mathbf{x})$  is equal to 1 if the documents  $\mathbf{x}$  is negative and zero if it is positive.

Once the regression parameters are learned with a Leave-One-Out cross validation, for each query the sizes of the positive set and the negative set are estimated with a cumulative approach as follows:

$$P_q = \alpha_P \cdot \mu_P(D_q) + \alpha_N \cdot \mu_N(D_q) \quad (12)$$

$$N_q = \alpha'_P \cdot \mu_P(D_q) + \alpha'_N \cdot \mu_N(D_q) \quad (13)$$



where  $q \in Q$ , and  $D_q$  is the result set computed using data set  $\Omega$ . Finally, the percentage for each category  $c$  and for each query  $q$  are

$$P_q^{\%} = \frac{P_q}{P_q + N_q} \quad N_q^{\%} = \frac{N_q}{P_q + N_q}. \quad (14)$$

Thanks to the additivity property of  $\mu_c$  we also note that Equations 12 and 13 are indeed the sum of 10 and 11 respectively.

## 6 Experiments

The evaluation measures for quantification models are based on a value aggregating the pairs with the observed and the predicted values for each category  $\{(y, \hat{y})\}_{c \in \mathcal{C}}$ . The main problem of sentiment quantification in a retrieval scenario is that there is a very high variability of sentiment category priors with real queries. Such a variance thus affects the performance of any classifier, and we therefore need to address specifically how to measure quantification performance with a set of queries and category observations,  $\{(y, \hat{y})\}_{c \in \mathcal{C}, q \in Q}$ .

However, the available collections for training and test the classifiers are of the order of few thousand of evaluated tweets, often on a single topic (OMD, HCR, GASP, WAB) or on a few similar topics (Sanders). SemEval contains many queries (about 180) but with very small validation and retrieval sets per query. There are about 3,000 tweets containing a 11% of negatives and 34% of positives with an average of 2 (5) negative (positive) tweets per query [19].

In addition, evaluation measures for quantification models have some drawbacks when applied to retrieval and quantification accuracy. One of evaluation measure used for quantification is the mean of the residuals (absolute errors, AE), that in our case would be the mean of the values  $y - \hat{y}$ , one for query and each category. AE is biased by the queries that possess a very large result set. To overcome this problem, one can use the mean of the rates (RAE) instead of the mean of absolute values. The bias with RAE is that it may be very low with very small category. To avoid the size of the result set, one can use the mean of divergence measure between the category distributions, e.g. the Kullback-Leibler divergence (KLD). Since KLD is unbounded, one can alternatively apply the logistic function to this divergence to normalize it, (NKLD) [7]. A comparison between models may be conducted with a statistical test, e.g. Wilcoxon, to show whether one model is statistically better than a second one[8].

Using a very different approach, statistical analysis studies the distribution of the residuals  $\{(y, \hat{y})\}_{c \in \mathcal{C}, q \in Q}$  with respect to their principal moments, in order to validate how much the learning model fits the data within a given confidence level (the margin of allowed error). The first advantage of using statistical analysis for quantification is that we can assess how good is a model independently from other learning models. Then we may always compare new models according only to the fitting parameters and their values. Second, the number of queries used to validate the fitting becomes an important parameter to pass the significance test of the goodness of the fit. Third, we can distinguish possible outliers of

the model, and we may correct hypothesis and improve models. For example, according to the normal Q-Q plot one should expect that all observations pairs should lie around a straight line. Therefore we use a validation collection [1] possessing the following properties:

- The collection should be large with a number of test queries  $Q$ .
- The result sets  $D_q$  of  $Q$  may vary largely from query to query.
- The size  $y = n_c^q$  of a category in each result set is estimated by  $\hat{y} = \hat{c}^q$  such that  $|\frac{n_c^q}{D_q} - \frac{\hat{c}^q}{D_q}| \leq \epsilon$  with a confidence level of 95%.

a) To assess the scalability of the cumulative hypothesis, we need to build a large collection of tweets using some generic terms as seeds and then running a number  $Q$  of queries. b) Differently from TREC collections where evaluation is focused on precision of rankings and it is thus performed by pooling the topmost (pseudo-relevant) documents from each running system, we here need to estimate the size of the set of relevant documents into categories. Therefore we estimate by sampling *randomly* from the result set, and assessing a *sufficient* (see condition c below) number  $V_q$  of documents with respect to six mutually exclusive categories: P for only positive, N for only Negative, X for neutral, M for mixed polarity, O for other, and the last category containing the rest of all non-relevant documents. All other five sentiment categories thus co-occur with the relevance event.

c) We then need to decide how many documents  $V_q$  to assess for each query in order to have a statistical significant estimate of category size within a given confidence interval. This confidence interval depends on the fixed confidence level, e.g. 95%. We know that the size of the validation set  $V_q$  must be of the

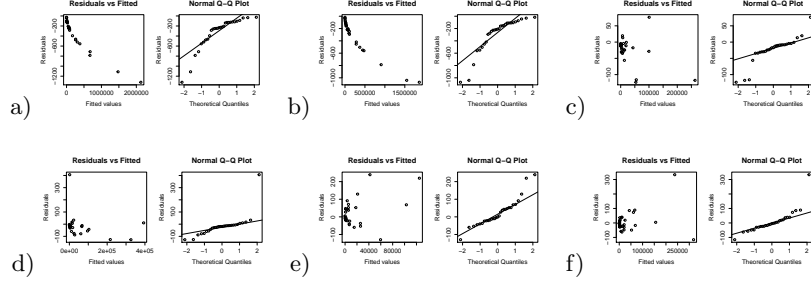
order of  $\frac{\sqrt{D_q}}{\sigma}$  where  $\sigma$  is the standard deviation of the category size.

d) To avoid the query bias, we use the Leave-One-Out cross-validation to learn the classifier without the query  $q \in Q$  and obtain  $\{(y, \hat{y})\}_{c \in C}$  of this query.

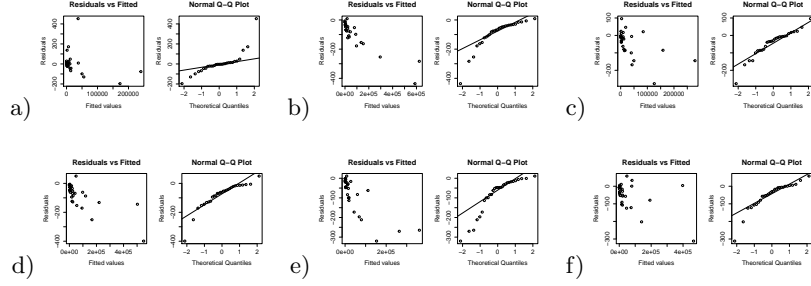
e) Finally, we study the distribution of the residuals to assess the precision of estimates for the categories. The collection contains 29 queries, with a median of 30,741 and a mean of 105,564 retrieved documents for a total of about 3 million retrieved documents. The rate of positive documents has mean 20.8%, the maximum 44.8% and the minimum 5.9%, while the mean rate of negative documents is 36.9%, the maximum 66.8% and the minimum 11.2%.

## 7 Conclusions

We have shown how to estimate sentiment categories proportions for retrieval through a non aggregative approach, and validating the approach with a very large result sets. The non aggregative approach is very efficient and suitable for real time analytics. The method consists in taking an additive measure  $\mu_c$  derived from the classifier and applied to the entire result set of a query in one single shot. We have also given the cumulative conditions under which such a measure can be defined for a given classifier. The model ignores the category priors, but



**Fig. 2.**  $\Phi$  with the **item driven** approach. **DBM**[a) b)], **MNB** [c) d)], **SVM** [e) f)]. Positive fitted values [a) c) e)], Negative fitted values [b) d) f)].



**Fig. 3.** Classify and Count. **DBM**[a) b)], **MNB** [c) d)], **SVM** [e) f)]. Positive fitted values [a) c) e)], Negative fitted values [b) d) f)].

to compensate this, it learns how to resize the cumulative measure through a linear regression model  $\Phi$ . We have used item-driven and query-driven settings for  $\Phi$ , and used three classifiers, two Multinomial Naive Bayes and SVM. We have used the Kolmogorov-Smirnov test to validate the hypothesis that observed and fitted values come from the same distribution for each method and classifier. In addition, we have used Pearson's correlation test to show that a linearity between the counting and the measures  $\mu_c$  is very strong. The results are also compared with the ACC baseline. Both item-driven and query driven approaches work similarly, but only SVM and MNB pass the Kolmogorov-Smirnov test for both positive and negative categories. There is the negative exception of DBM with the item-driven approach. The ACC confirms to be not stable or reliable both with Pearson correlation and Kolmogorov-Smirnov test.

## References

1. Giambattista Amati, Simone Angelini, Marco Bianchi, Luca Costantini, and Giuseppe Marcone. A scalable approach to near real-time sentiment analysis on social networks. In CEUR-WS.org, editor, *DART, Proceedings of the 8th International Workshop on Information Filtering and Retrieval*, volume 1314, pages 12–23, 2014.
2. José Barranquero, Jorge Díez, and Juan José del Coz. Quantification-oriented learning based on reliable classifiers. *Pattern Recognition*, 48(2):591–604, 2015.
3. Antonio Bella, César Ferri, José Hernández-Orallo, and M. José Ramírez-Quintana. Aggregative quantification for regression. *Data Min. Knowl. Discov.*, 28(2):475–518, 2014.
4. Nicholas A. Diakopoulos and David A. Shamma. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1195–1198, New York, NY, USA, 2010. ACM.
5. George Forman. Counting positives accurately despite inaccurate classification. In João Gama, Rui Camacho, Pavel Brazdil, Alípio Jorge, and Luís Torgo, editors, *ECML*, volume 3720 of *Lecture Notes in Computer Science*, pages 564–575. Springer, 2005.
6. George Forman. Quantifying counts and costs via classification. *Data Min. Knowl. Discov.*, 17(2):164–206, 2008.
7. Wei Gao and Fabrizio Sebastiani. Tweet sentiment: From classification to quantification. In IEEE/ACM, editor, *ASONAM, International Conference on Advances in Social Networks Analysis and Mining*, 2015.
8. Víctor González-Castro, Rocío Alaiz-Rodríguez, and Enrique Alegre. Class distribution estimation based on the hellinger distance. *Inf. Sci.*, 218:146–164, January 2013.
9. Daniel Hopkins and Gary King. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247, 01/2010 2010.
10. Thorsten Joachims. A support vector method for multivariate performance measures. In Luc De Raedt and Stefan Wrobel, editors, *Proceedings of the 22nd International Conference on Machine Learning (ICML 2005), August 7-11, 2005, Bonn, Germany*, pages 377–384. ACM Press, New York, NY, USA, 2005.

11. Solomon Kullback. *Information Theory and Statistics*. Wiley, New York, 1959.
12. Patrice Latinne, Marco Saerens, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities may significantly improve classification accuracy: Evidence from a multi-class problem in remote sensing. *NEURAL COMPUTATION*, 14:14–21, 2001.
13. D. C. Van Leijenhorst and Theo P. Van Der Weide. A formal derivation of Heaps’ law. *Inf. Sci.*, 170(2-4):263–272, 2005.
14. P S Levy and E H Kass. A three-population model for sequential screening for bacteriuria. *American J. of Epidemiology*, 91(2):148–54, 1970.
15. Benoit Mandelbrot. On the theory of word frequencies and on related markovian models of discourse. In *Proceedings of Symposia in Applied Mathematics. Vol. XII: Structure of language and its mathematical aspects*, pages 190–219. American Mathematical Society, Providence, R.I., 1961. Roman Jakobson, editor.
16. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.
17. Jason D. Rennie, Lawrence Shih, Jaime Teevan, and David R. Karger. Tackling the poor assumptions of naive bayes text classifiers. In *ICML*, pages 616–623. AAAI Press, 2003.
18. A. Renyi. *Foundations of probability*. Holden-Day Press, San Francisco, USA, 1969.
19. Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 451–463, Denver, Colorado, June 2015. Association for Computational Linguistics.
20. Jack Chongjie Xue and Gary M. Weiss. Quantification and semi-supervised classification methods for handling changes in class distribution. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’09, pages 897–906, New York, NY, USA, 2009. ACM.